Learning Acoustic Word Embeddings from Sequence-to-Sequence Models

Shruti Palaskar
What is this talk about?

How to cram meaning of speech into a vector!?!
"You can't cram the meaning of a whole %&!$# sentence into a single $&!#* vector!"

- Raymond Mooney
How to *try to* cram the meaning of a whole sentence into a single vector?

➢ ELMo, BERT
➢ word2vec, glove
Text Embeddings

➢ Representing written words or sentences as continuous valued fixed dimensional vectors
➢ Common representation for various words/sentences/languages
➢ Useful as off-the-shelf pre-trained features for other tasks
Acoustic Embeddings

➢ Map speech signal of arbitrary length into a fixed dimensional vector
➢ This speech signal may be for a word or a sentence

[Figure credit: Herman Kamper]
Acoustic Embeddings

➢ Represent speech (an inherently continuous signal) into embeddings (fixed dimensional vectors)

➢ Speech has many more variations than text like:
  - speaking rate, pronunciation variance, speaker differences,
  - acoustic environment, prosody (emotion etc), intonation, ...

➢ Can we do the same with speech as text then? Let's see...
Acoustic Embedding: Uses & Applications

➢ Speech Similarity tasks
  ○ Spoken Language Understanding
  ○ Whole-word Speech Recognition
  ○ Spoken Term Discovery
  ○ Query-by-example

[Figure credit: Herman Kamper]
Acoustic Embedding: Uses & Applications

➢ Shared representation for speech and other modalities (like text or vision)
  ○ Easier multimodal interaction for these different modalities
  ○ Given speech, retrieve text / Given speech retrieve corresponding video!

Speech segment of “CAT”

CAT
Talk Outline

I. Learning Acoustic Word Embeddings
   A. Model: Acoustic-to-Word Speech Recognition
   B. Understanding A2W models
   C. Evaluation

II. Applications of Acoustic Word Embeddings
   A. Spoken Language Understanding
   B. Unsupervised speech recognition and spoken language translation
I. Learning Acoustic Word Embeddings
   A. Model: Acoustic-to-Word Speech Recognition
   B. Understanding A2W models
   C. Evaluation

II. Applications of Acoustic Word Embeddings
   A. Spoken Language Understanding
   B. Unsupervised speech recognition and spoken language translation
Acoustic-to-Word Speech Recognition

This Speech Recognizer can *Recognize Speech*
Acoustic-to-Word Speech Recognition

This Speech Recognizer can *Wreck a Nice Beach*
Acoustic-to-Word Speech Recognition

- Model Probability (Words | Acoustics)
- These acoustics could be *any* form of representation of speech
- Sequence-to-Sequence model with attention
- Around 30,000 words vocabulary
- Usually 26 character vocabulary (English)
- No alignment needed like traditional speech recognizers

Chan et al., "Listen, Attend and Spell", 2016
Results

This Speech Recognizer can Wreck a Nice Beach

➢ Evaluation: Word Error Rate
➢ On a standard dataset Switchboard

Character models = 15.6%
Word models = 22.1%

➢ But whole words are semantically meaningful units!
➢ Can perform non-speech transcription task with speech input!

Palaskar and Metze, “Acoustic-to-Word Recognition with Sequence-to-Sequence Models”, 2018
Talk Outline

I. Learning Acoustic Word Embeddings
   A. Model: Acoustic-to-Word Speech Recognition
   B. Understanding A2W models
   C. Evaluation

II. Applications of Acoustic Word Embeddings
   A. Spoken Language Understanding
   B. Unsupervised speech recognition and spoken language translation
Understanding Acoustic-to-Word Models

This Speech Recognizer can Wreck a Nice Beach
Location-aware Attention

➢ Attention is a rich source of interpretability and understanding in sequence-to-sequence models

➢ Specially, input speech and output text are monotonic signals unlike Machine Translation or summarization

➢ Monotonicity: time-synchronous alignments only

➢ Can enforcing monotonicity help improve ASR performance? Yes. [Chan et al., “Listen, attend and spell”, 2015]

➢ New attention mechanism for sequence-to-sequence based ASR
Analyzing Attention

➢ Each color corresponds to a word in the output
➢ Highly localized attention
➢ Peaky distribution
➢ Last word attention is non-peaky
➢ Time steps 80-100 are silence in speech

Palaskar and Metze, “Acoustic-to-Word Recognition with Sequence-to-Sequence Models”, 2018
What is the model learning?

➢ Q1. What does it mean that attention is peaky/localized for a word?

➢ Model focuses on a single input speech frame for every word

➢ Model localizes word boundaries without supervision

Palaskar and Metze, “Acoustic-to-Word Recognition with Sequence-to-Sequence Models”, 2018
What is the model learning?

➢ Q2. What does it mean that attention is “absent” between timesteps 80-100?

➢ Model learns to detect speech and non-speech segments without supervision

Palaskar and Metze, “Acoustic-to-Word Recognition with Sequence-to-Sequence Models”, 2018
Q3. What does every peak corresponding to a word represent?

➢ It represents a single fixed-size representation of input speech, or the **acoustic word embedding**

Palaskar and Metze, “Acoustic-to-Word Recognition with Sequence-to-Sequence Models”, 2018
What *all* is the model learning?

1. The model focuses on a single input speech frame for every word
2. It localizes word boundaries in continuous speech without supervision
3. It learns to detect speech and non-speech segments in continuous speech without supervision
4. It represents every output word as a single fixed-size representation of input speech, or the *acoustic word embedding*

Palaskar and Metze, “Acoustic-to-Word Recognition with Sequence-to-Sequence Models”. 2018

Shruti Palaskar
Carnegie Mellon University
Learning Contextual Acoustic Word Embeddings

Palaskar*, Raunak* and Metze, "Learned in Speech Recognition: Contextual Acoustic Word Embeddings", 2019


Shruti Palaskar

Carnegie Mellon University
Using Attention to learn CAWE

\[ w_i = \frac{\sum_{k \in K} \text{encoder}(a_k)}{n(K)} \]  

\[ w_i = \frac{\sum_{k \in K} \text{attention}(a_k) \cdot \text{encoder}(a_k)}{n(K)} \]  

\[ w_i = \text{encoder}(a_k) \text{ where } k = \arg \max_{k \in K} \text{attention}(a_k) \]

(1) U-AVG: Averaged without attention weights  
(2) CAWE-W: Averaged with attention weights  
(3) CAWE-M: Arg max of attention weights

➢ Choose based on application

Talk Outline

I. Learning Acoustic Word Embeddings
   A. Model: Acoustic-to-Word Speech Recognition
   B. Understanding A2W models
   C. Evaluation

II. Applications of Acoustic Word Embeddings
   A. Spoken Language Understanding
   B. Unsupervised speech recognition and spoken language translation
Evaluating Acoustic Word Embeddings

➢ Standard Sentence Embedding Evaluation Benchmarks

➢ There are 17 standard sentence evaluation benchmarks in NLP

➢ Most new methods to evaluate sentence embeddings are scored on these methods for fair evaluation

➢ We compare CAWE with text-based word2vec embeddings learned on the transcripts

➢ A2W models trained on Switchboard (conversational) and How2 (planned but free speech, outdoors, distance microphone)


Shruti Palaskar
Carnegie Mellon University
SentEval

➢ Standard Sentence Embedding Evaluation Benchmarks
➢ Fixed datasets on Sentence Textual Similarity, classification (movie reviews, product reviews etc), entailment, sentiment analysis, question type etc.

➢ **Human annotated** similarity scores present for this dataset
➢ Proposed *word* embeddings are plugged for all words in a *sentence* (1)
➢ Similarly, baseline *word* embeddings are plugged in for all words in a *sentence* (2)
➢ Correlation or Classification scores are computed with these two *sentence* embeddings

https://github.com/facebookresearch/SentEval
Comparing CAWE methods

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Switchboard</th>
<th>How2</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>U-AVG</td>
<td>CAWE-W</td>
</tr>
<tr>
<td>STS 2012</td>
<td>0.3230</td>
<td>0.3281</td>
</tr>
<tr>
<td>STS 2013</td>
<td>0.1252</td>
<td>0.1344</td>
</tr>
<tr>
<td>STS 2014</td>
<td>0.3358</td>
<td>0.3389</td>
</tr>
<tr>
<td>STS 2015</td>
<td>0.3854</td>
<td>0.3881</td>
</tr>
<tr>
<td>STS 2016</td>
<td>0.2998</td>
<td>0.2974</td>
</tr>
<tr>
<td>STS B</td>
<td>0.3667</td>
<td>0.3510</td>
</tr>
<tr>
<td>SICK-R</td>
<td>0.5640</td>
<td>0.5800</td>
</tr>
</tbody>
</table>

- CAWE-M always performs better in STS tasks
- CAWE-W more generalizable but noisy
- U-AVG noisiest

Comparing CAWE with word2vec

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Switchboard</th>
<th></th>
<th>How2</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>CAWE-M</td>
<td>CBOW</td>
<td>Concat</td>
<td>CAWE-M</td>
<td>CBOW</td>
</tr>
<tr>
<td>STS 2012</td>
<td>0.3561</td>
<td>0.3639</td>
<td>0.3470</td>
<td>0.3648</td>
<td>0.3688</td>
</tr>
<tr>
<td>STS 2013</td>
<td>0.1969</td>
<td>0.1960</td>
<td>0.2010</td>
<td>0.2716</td>
<td>0.2524</td>
</tr>
<tr>
<td>STS 2014</td>
<td>0.3888</td>
<td>0.3745</td>
<td>0.3795</td>
<td>0.3940</td>
<td>0.3973</td>
</tr>
<tr>
<td>STS 2015</td>
<td>0.4275</td>
<td>0.4459</td>
<td>0.4481</td>
<td>0.4173</td>
<td>0.4781</td>
</tr>
<tr>
<td>STS 2016</td>
<td>0.3833</td>
<td>0.3471</td>
<td>0.3651</td>
<td>0.3159</td>
<td>0.4023</td>
</tr>
<tr>
<td>STS B</td>
<td>0.401</td>
<td>0.4100</td>
<td>0.3995</td>
<td>0.4000</td>
<td>0.4720</td>
</tr>
<tr>
<td>SICK-R</td>
<td>0.6006</td>
<td>0.6170</td>
<td>0.6228</td>
<td>0.6440</td>
<td>0.6550</td>
</tr>
</tbody>
</table>

CAWE performs competitively with word2vec

Improvement in concatenation shows both embeddings contribute unique features

Gains more prominent in SWBD as it is conversational while How2 is planned
Talk Outline

I. Learning Acoustic Word Embeddings
   A. Model: Acoustic-to-Word Speech Recognition
   B. Understanding A2W models
   C. Evaluation

II. Applications of Acoustic Word Embeddings
   A. Spoken Language Understanding
   B. Unsupervised speech recognition and spoken language translation
Spoken Language Understanding

➢ Speech-based downstream task other than transcription
➢ ATIS dataset of flight queries with intent, domain, and named entities
➢ Widely used corpus for SLU
➢ Classification Task: Given query identify intent, domain and named entities
➢ Prior work used transcription of speech rather than audio input for this task [Mesnil et al. 2013]
➢ Performance in this task will help validate use of CAWE
Using CAWE for Spoken Language Understanding

<table>
<thead>
<tr>
<th></th>
<th>CAWE-M</th>
<th>CAWE-W</th>
<th>CBOW</th>
</tr>
</thead>
<tbody>
<tr>
<td>RNN</td>
<td>91.49</td>
<td>91.67</td>
<td>91.82</td>
</tr>
<tr>
<td>GRU</td>
<td>93.25</td>
<td>93.56</td>
<td>93.11</td>
</tr>
</tbody>
</table>

➢ Two simple models: RNN and GRU
➢ F1 score for classification on CAWE-M, CAWE-W and CBOW
➢ CAWE performs competitively with text embeddings highlighting its utility
➢ Can be used as off-the-shelf embeddings for other speech-based tasks when trained on larger data

Talk Outline

I. Learning Acoustic Word Embeddings
   A. Model: Acoustic-to-Word Speech Recognition
   B. Understanding A2W models
   C. Evaluation

II. Applications of Acoustic Word Embeddings
   A. Spoken Language Understanding
   B. Unsupervised speech recognition and spoken language translation
Once I have my jack stand there on the rear axle, go ahead and release the hydraulic pressure...

Quando eu tiver meu macaco parado no eixo traseiro, vá em frente e libere a pressão hidráulica...

Changing flat tires doesn’t have to be done with car jacks. Learn how to use an automotive hydraulic lift...

Sanabria et al., "How2: A Large Scale dataset for Multimodal Language Understanding", 2018
The big picture

So as you can see I added some sesame seed, some black sesame seed here in my plate.

Subtitle

Translation

Como vocês podem ver, eu coloquei no meu prato o gergelim preto.

Transcription

So as you can see I added some sesame seed, some black sesame seed here in my plate.

Summary

A cooking recipe for Seared Sesame Crusted Tuna with Wild Rice.
Learning Multimodal Embeddings

I. Each is different but all views share similar information

II. Visual, Auditory and Language views are aligned

III. Views in the same modality v/s Views in multiple modalities

IV. Unit level representations v/s Sequence Level Representations


Shruti Palaskar
Carnegie Mellon University
Canonical Correlation Analysis

Task Specific Representations

Transformations

Correlated Cross View Semantic Space

Concept I

Concept E

Concept P

Shruti Palaskar

Carnegie Mellon University
CCA in a Nutshell

Pairs of points: \((X, Y) \sim \mathcal{D}_{X,Y}\)

View 1  View 2

“A man in an orange hat staring at something.”

Find transformations \(\mathbf{u} \in \mathbb{R}^{d_x}, \mathbf{v} \in \mathbb{R}^{d_y}\)

to maximize \(\text{correlation}(\mathbf{u}^T f_\theta(X), \mathbf{v}^T g_\phi(Y))\)

Hotelling, 1936; Wang et al., 2016
Text Representations - Sentences

Encoder trained for MT

2-layer BiGRU

mean pool

English Text

Shruti Palaskar
Carnegie Mellon University
Video Representations

“Bag-of-classes” representation

ResNet multi-class posterior

ResNet
ResNet
ResNet
ResNet

meanpool
Speech Representations - Sentences [CAWE]

Palaskar and Metze, “Acoustic-to-Word Recognition with Sequence-to-Sequence Models”, 2018
Speech and Text Representations

Retrieval of Text Given Speech

CCA

Recall@10 over Test set

Linear CCA: 96.9%
Deep CCA: 90.1%

Holzenberger*, Palaskar*, Madhyastha, Metze and Arora., "Learning from Multiview Correlations in Open-Domain Videos", 2019
Retrieve Speech Given Text

CCA

Recall@10 over Test set

Linear CCA
96.1%

Deep CCA
89.7%

Holzenberger*, Palaskar*, Madhyastha, Metze and Arora., "Learning from Multiview Correlations in Open-Domain Videos", 2019
Speech and Video Representations

Holzenberger*, Palaskar*, Madhyastha, Metze and Arora., "Learning from Multiview Correlations in Open-Domain Videos", 2019
Retrieve Video Given Speech

Speech, Text and Video Representations

Holzenberger*, Palaskar*, Madhyastha, Metze and Arora., "Learning from Multiview Correlations in Open-Domain Videos", 2019
<table>
<thead>
<tr>
<th></th>
<th>English Text</th>
<th>Portuguese Text</th>
</tr>
</thead>
<tbody>
<tr>
<td>-</td>
<td>85.4</td>
<td>70.7</td>
</tr>
<tr>
<td>85.4</td>
<td>-</td>
<td>98.4</td>
</tr>
<tr>
<td>71.0</td>
<td>98.3</td>
<td>-</td>
</tr>
<tr>
<td>1.1</td>
<td>1.1</td>
<td>0.9</td>
</tr>
</tbody>
</table>

## Retrieve Text Given Speech - Comparison

<table>
<thead>
<tr>
<th>Model</th>
<th>Recall@10</th>
</tr>
</thead>
<tbody>
<tr>
<td>Speech &amp; En Text</td>
<td>90.1%</td>
</tr>
<tr>
<td>Speech, En Text, Pt Text &amp; Video</td>
<td>85.4%</td>
</tr>
</tbody>
</table>

Retrieval for ASR

Given a Speech segment from the test set, retrieve the closest English sentence in a reference set.

<table>
<thead>
<tr>
<th>Reference set</th>
<th>WER</th>
</tr>
</thead>
<tbody>
<tr>
<td>S2S Model</td>
<td>24.2 %</td>
</tr>
<tr>
<td>Train</td>
<td>134 %</td>
</tr>
<tr>
<td>Train + Test</td>
<td>27.4 %</td>
</tr>
</tbody>
</table>

Holzenberger*, Palaskar*, Madhyastha, Metze and Arora., "Learning from Multiview Correlations in Open-Domain Videos", 2019
Retrieval for SLT

Given a Speech segment from the test set, retrieve the closest Portuguese sentence in a reference set.

<table>
<thead>
<tr>
<th>Reference set</th>
<th>BLEU</th>
</tr>
</thead>
<tbody>
<tr>
<td>S2S Model</td>
<td>27.9</td>
</tr>
<tr>
<td>Train</td>
<td>0.2</td>
</tr>
<tr>
<td>Train + Test</td>
<td>19.8</td>
</tr>
</tbody>
</table>

Holzenberger*, Palaskar*, Madhyastha, Metze and Arora., "Learning from Multiview Correlations in Open-Domain Videos", 2019
To conclude
Main Takeaways

1. Possible to learn pre-trained acoustic word embeddings similar to text (bert, elmo) and vision (alexnet, vggnet)

2. These embeddings perform well with text based embeddings and capture complimentary information than text embeddings

3. Can perform non-transcription tasks with speech inputs: spoken language understanding

4. Can learn shared global multimodal embedding spaces to perform unsupervised ASR, SLT etc
Main Takeaways

1. Possible to learn pre-trained acoustic word embeddings similar to text (bert, elmo) and vision (alexnet, vggnet)

2. AWE performs competitively with word2vec and capture complimentary information than text embeddings

3. Can perform non-transcription tasks with speech inputs: spoken language understanding

4. Can learn shared global multimodal embedding spaces to perform unsupervised ASR, SLT etc
Main Takeaways

1. Possible to learn pre-trained acoustic word embeddings similar to text (bert, elmo) and vision (alexnet, vggnet)

2. These embeddings perform well with text based embeddings and capture complimentary information than text embeddings

3. Can perform non-transcription tasks with speech inputs: spoken language understanding

4. Can learn shared global multimodal embedding spaces to perform unsupervised ASR, SLT etc
Main Takeaways

1. Possible to learn pre-trained acoustic word embeddings similar to text (bert, elmo) and vision (alexnet, vggnet)

2. These embeddings perform well with text based embeddings and capture complimentary information than text embeddings

3. Can perform non-transcription tasks with speech inputs: spoken language understanding

4. Can learn shared global multimodal embedding spaces to perform unsupervised ASR, SLT etc
Thank you!

Questions?

spalaska@cs.cmu.edu